Is This Translation Error Critical? Classification-Based Human and Automatic MT Evaluation Focusing on Critical Errors

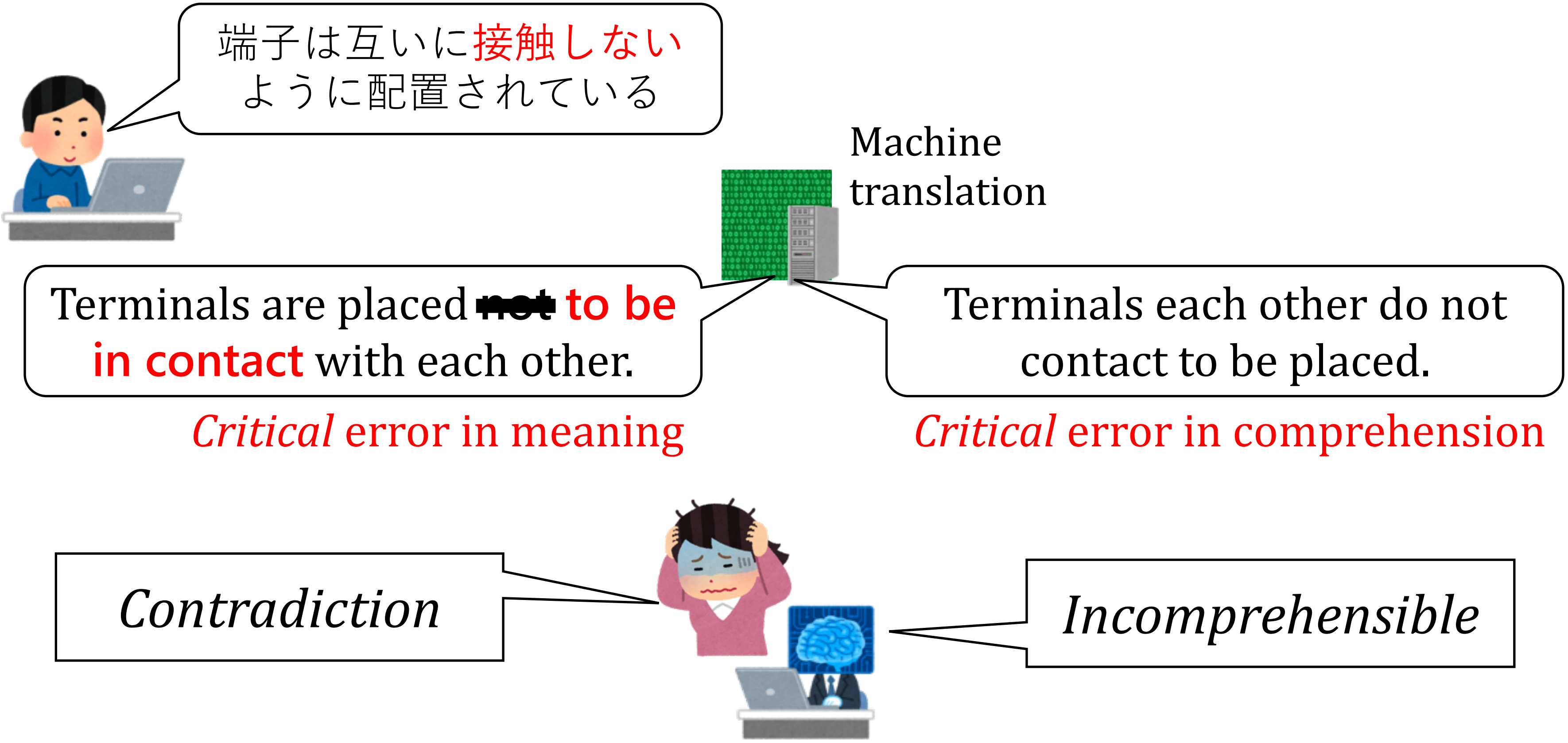
Katsuhito Sudoh*, Kosuke Takahashi, Satoshi Nakamura

Nara Institute of Science and Technology (NAIST)

*PRESTO, Japan Science and Technology Agency

This work is supported by JST PRESTO (Grant number JPMJPR1856).

Quick Overview



端子は互いに接触しないように配置されている

Machine translation

Terminals are placed ~~not~~ to be in contact with each other.

Critical error in meaning

Terminals each other do not contact to be placed.

Critical error in comprehension

Contradiction

Incomprehensible

Revisitng classification-based MT evaluation in two dimensions: Adequacy & Fluency

Background

- Regression-based MT evaluation
  - Founded on Human Direct Assessment (Graham+ 2016)
  - Predict human DA scores using reference and hypothesis translations
    - BERT regressor (Shimanaka+ 2019), BERTScore (Zhang+ 2020), BLEURT (Sellam+ 2020), …
- Can they identify critial errors?

(Artificial) Examples

Examples	BLEU	BERT Score	BLEURT
The Pleiades is situated 445 light-years from Earth. [same as ref.]	1.00	1.00	0.94
The Pleiades is not situated 445 light-years from Earth.	0.70	0.89	0.03
The Pleiades is situated 445 light-years from Mars.	0.78	0.91	0.64
Is Earth from Pleiades the light-years situated cluster 445.	0.07	0.49	-0.66
Turn off the light for saving the Earth.	0.09	0.04	-1.55

Research objective

- Classification-based MT evaluation that can identify such critical errors
    - Two-dimensional
        - Fluency (including comprehension)
        - Adequacy
    - Sentence-based
        - cf. segment-level annotations by Popovic (CoNLL and COLING 2020)
- Both human and automatic evaluation

Human evaluation

- Dataset: WMT Metrics Task (2015-17)
  - 9,280 MT results in English
- A linguistic data development company hired *three* annotators:
  - Native speakers of English
  - Work experience in translation into English
  - No specific training conducted

Human evaluation (cont'd)

- Evaluation in a *monolingual* way
    - The annotators can see only MT results along with corresponding references
- Independent among the annotators
- The evaluation corpus is available under CC BY-NC-SA 4.0
    - https://github.com/ksudoh/wmt15-17-humaneval

Evaluation criteria

Fluency

Incomprehensible (F)	The sentence is not comprehensible.
Poor (D)	Some contents are not easy to understand by typographical / grammatical errors and problematic expressions.
Fair (B)	All the contents are easy to understand in spite of some typographical / grammatical errors.
Good (A)	All the contents are easy to understand and free from grammatical errors, but some expressions are not very fluent.
Excellent (S)	All the contents are easy to understand, and all the expressions are flawless.

Adequacy

Incomprehensible (F)	The contents cannot be understood due to fluency and comprehension issues, so the hypothesis is not eligible for the adequacy evaluation.
Unrelated (O)	The hypothesis delivers information that is *not related* to the reference
Contradiction (C)	The hypothesis delivers information that *contradicts* the reference
Serious (F)	The hypothesis delivers information that may cause serious misunderstanding due to some content errors but does not contradict the reference
Fair (B)	All the contents are easy to understand in spite of some typographical / grammatical errors.
Good (A)	All the contents are easy to understand and free from grammatical errors, but some expressions are not very fluent.
Excellent (S)	All the contents are easy to understand, and all the expressions are flawless.

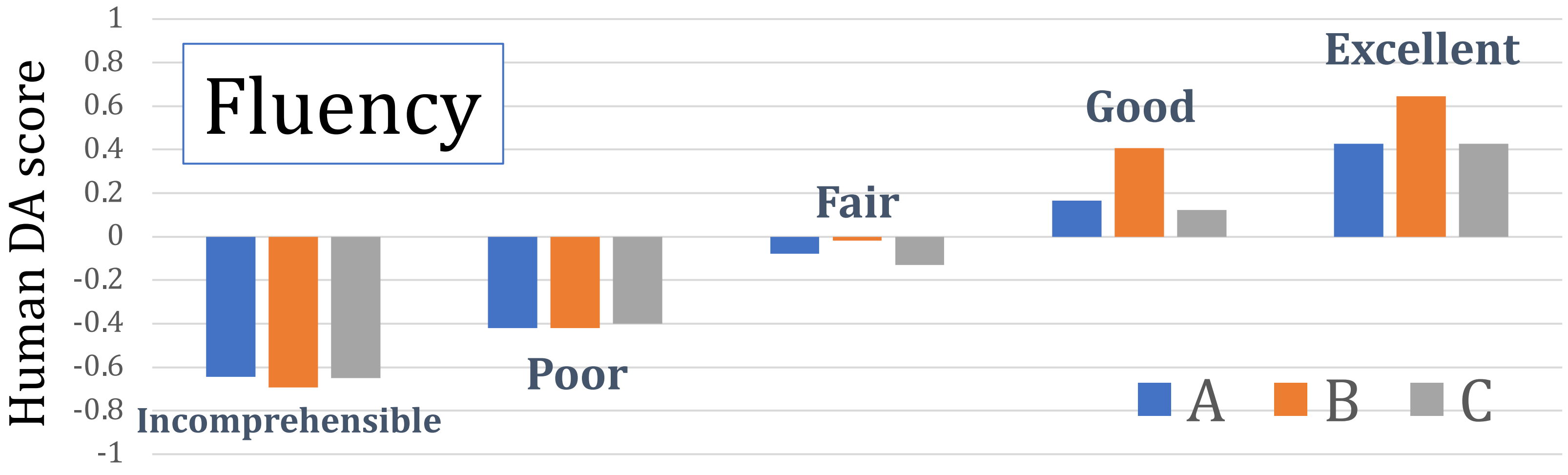
Analysis: Agreement

- Agreement was not high (~0.3)
  - Similar to previous studies with older WMT datasets (Callison-Burch+ 2007)
  - Annotator B was strict in Fluency
  - Annotator C was strict in Adequacy

		A-B	A-C	B-C
Fluency	Kappa	.286	.377	.249
	Concordance	.451	.511	.401
Adequacy	Kappa	.395	.268	.277
	Concordance	.546	.587	.575

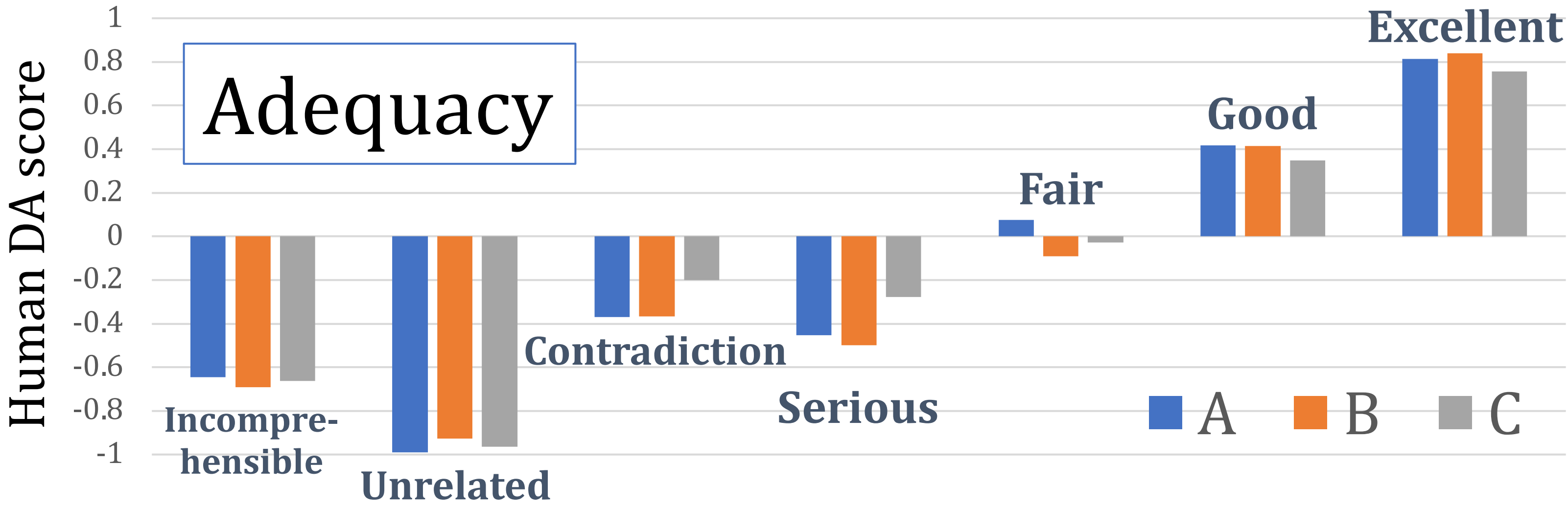
Analysis: Human DA scores

- Fluency almost works as a Likert scale



Analysis: Human DA scores

- Interesting finding in Adequacy
  - *Unrelated* hypotheses are scored worst
  - *Contradiction* are scored better than other error categories (due to the similarity with the reference contents?)



Automatic evaluation

- A RoBERTa-based classifier model
- Data split
  - Training: 4,824 from 2015-16
  - Development: 536 from 2015-16
  - Test: 3,920 from 2017 (560 each for {cs,de,fi,lv,ru,tr,zh}-en)
- Label agreement among annotators
  - Majority
  - Pessimistic heuristics (details in paper)

Results by confusion matrix

- Fluency accuracy: 57.8%
  - Serious confusion between adjacent categories

>> Prediction

>> Correct labels

Fluency	Inc.	Poor	Fair	Good	Exc.
Incomprehensible	206	45	22	8	1
Poor	45	266	250	43	4
Fair	15	134	782	358	52
Good	2	11	187	560	139
Excellent	0	2	35	306	453

Results by confusion matrix

- Adequacy accuracy: 60.0%
  - Confusion between Serious and Fair

Adequacy	Inc.	Unr.	Con.	Ser.	Fair	Good	Exc.
Incomprehensible	224	0	0	83	38	4	1
Unrelated	0	1	0	13	5	0	0
Contradiction	0	0	8	9	13	10	0
Serious	37	0	8	385	242	45	0
Fair	29	0	13	237	878	274	10
Good	4	0	9	20	302	77	59
Excellent	0	0	0	1	6	97	84

Summary of the results

- Fluency
  - # of serious classification errors with distant categories was small
- Adequacy
  - Less frequent categories (*Unrelated* and *Contradiction*) were difficult to predict
  - Prediction of *Excellent* seemed good; their actual judgements were mostly *Excellent* or *Good* (93.5%)

Conclusions

- Classification-based human and automatic MT evaluation
  - Fluency & Adequacy, motivated
- Human evaluation can be improved for better agreement
  - More careful evaluation instruction?
- Automatic evaluation should be improved for the practical use

Future work

- Further development of human evaluation corpora, not limited to WMT Metrics Task
- Data augmentation to tackle the label imbalance
  - Shared task data does not fully cover actual MT problems…
- MT training/fine-tuning based on these evaluation criteria